An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

Ryan Giordano
MIT

Rachael Meager
LSE

Tamara Broderick
MIT

Job talk 2021

Dropping data: Motivation

You're a data analyst, and you've

- Gathered some exchangeable data,
- Cleaned up / removed outliers,
- Checked for correct specification, and
- Drawn a conclusion from your statistical analysis
  (e.g., based the sign / significance of some estimated parameter).

Dropping data: Motivation

You're a data analyst, and you've

- Gathered some exchangeable data,
- Cleaned up / removed outliers,
- Checked for correct specification, and
- Drawn a conclusion from your statistical analysis (e.g., based the sign / significance of some estimated parameter).

Well done!

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!

Dropping data: Mexico Microcredit

Consider Angelucci et al. [2015], a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points. The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!
Question: Is the reported interval $-4.55 \pm (5.88)$ a reasonable description of the uncertainty in the estimated efficacy of microcredit?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?
Not always!

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Not always!

...but sometimes, surely yes.

For example, often in economics:

- Small fractions of data are missing not-at-random,
- Policy population is different from analyzed population,
- We report a convenient summary (e.g. mean) of a complex effect,
- Models are stylized proxies of reality.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?
The number of subsets $\binom{N}{\lfloor \alpha N \rfloor}$ can be very large even when α is very small.
In the MX microcredit study, $\binom{16560}{15} \approx 1.4 \cdot 10^{51}$ sets to check for $\alpha = 0.0009$.
We provide a fast, automatic approximation based on the **influence function**.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Non-robustness to removal of $\lfloor \alpha N \rfloor$ points is:

- Not (necessarily) caused by misspecification.
- Not (necessarily) caused by outliers.
- Not captured by standard errors.
- Not mitigated by large N.
- Primarily determined by the **signal to noise** ratio

  ... in a sense which we will define.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

- We provide deterministic error bounds for small α.
- We show the accuracy in simple experiments.
- We show the accuracy in a number of real-world experiments.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Conclusion: Related work and future directions

Question 1:
How do we find influential datapoints?

Question 2:
What makes an estimator non-robust?

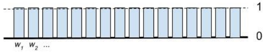
Question 3:
When is our approximation accurate?

The influence function

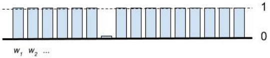
- Weights as derivatives
- Influence function
- Simulation
- Experiments

The linear approximation.

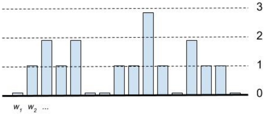
Original weights:

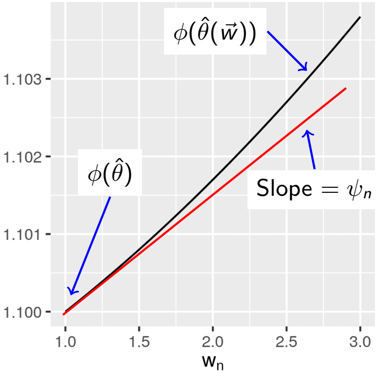


Leave-one-out weights:



Bootstrap weights:





$$\phi(\hat{\theta}(\vec{w})) = \phi(\hat{\theta}) + \sum_{n=1}^{N} \psi_n(\vec{w}_n - 1) + \text{Higher-order derivatives}$$

Key idea: Controlling higher-order derivatives can control the error.

The linear approximation.

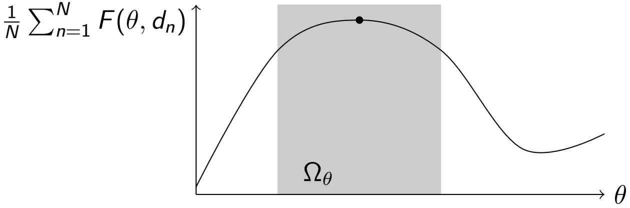
Let W_α be the set of weight vectors with no more than $\lfloor \alpha N \rfloor$ zeros.

Let $H(\theta, d_n) := \left. \frac{\partial G(\theta, d_n)}{\partial \theta^T} \right|_\theta$.

Assumption (Smooth Objective)

Fix the dataset. Assume there exists a compact $\Omega_\theta \subseteq \mathbb{R}^D$ with $\hat{\theta}(\vec{w}) \in \Omega_\theta$ for all $\vec{w} \in W_\alpha$. Assume that, for all $\theta \in \Omega_\theta$:

- $\frac{1}{N} \sum_{n=1}^N H(\theta, d_n)$ *and* $\frac{1}{N} \sum_{n=1}^N G(\theta, d_n)$ *are bounded.*
- $\frac{1}{N} \sum_{n=1}^N H(\theta, d_n)$ *is uniformly non-singular and Lipschitz (in θ).*
- $\phi(\theta)$ *has a Lipschitz first derivative.*



The linear approximation.

Theorem

Let Assumption 1 hold for a given dataset. Then there exists a sufficiently small α *such that*

$$\sup_{\vec{w}\in W_\alpha}\left|\phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}(\vec{w}))\right| \le C_1\alpha \; and \; \sup_{\vec{w}\in W_\alpha}\left|\phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta})\right| \le C_2\sqrt{\alpha},$$

where C_1 *and* C_2 *are given by the quantities in the assumption.*

The linear approximation.

Theorem

Let Assumption 1 hold for a given dataset. Then there exists a sufficiently small α *such that*

$$\sup_{\vec{w} \in W_\alpha} \left| \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}(\vec{w})) \right| \leq C_1 \alpha \ \ and \ \ \sup_{\vec{w} \in W_\alpha} \left| \phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \right| \leq C_2 \sqrt{\alpha},$$

where C_1 *and* C_2 *are given by the quantities in the assumption.*

Since $\alpha \ll \sqrt{\alpha}$ when α is small, Theorem 1 states that the linear approximation's error is of smaller order than the actual difference.

The linear approximation.

Theorem

Let Assumption 1 hold for a given dataset. Then there exists a sufficiently small α *such that*

$$\sup_{\vec{w} \in W_\alpha} \left| \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}(\vec{w})) \right| \leq C_1 \alpha \text{ and } \sup_{\vec{w} \in W_\alpha} \left| \phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \right| \leq C_2 \sqrt{\alpha},$$

where C_1 *and* C_2 *are given by the quantities in the assumption.*

Proof sketch.

The second inequality follows from the smoothness of the objective. The first inequality follows from the smoothness of $d\hat{\theta}(\vec{w})/d\vec{w}$. □

The linear approximation.

Theorem

Let Assumption 1 hold for a given dataset. Then there exists a sufficiently small α such that

$$\sup_{\vec{w}\in W_\alpha} \left| \phi^{\mathrm{lin}}(\vec{w}) - \phi(\hat{\theta}(\vec{w})) \right| \le C_1 \alpha \textit{ and } \sup_{\vec{w}\in W_\alpha} \left| \phi(\hat{\theta}(\vec{w})) - \phi(\hat{\theta}) \right| \le C_2 \sqrt{\alpha},$$

where C_1 and C_2 are given by the quantities in the assumption.

Proof sketch.

The second inequality follows from the smoothness of the objective. The first inequality follows from the smoothness of $d\hat{\theta}(\vec{w})/d\vec{w}$. □

Corollary

Under standard conditions, Assumption 1 holds for fixed constants with probability approaching one for $N \to \infty$. Then Theorem 1 applies with probability approaching one as $N \to \infty$.

The linear approximation.

For $N = 5,000$ data points, compute the OLS estimator from:

Regressors	Residuals	Responses
$x_n \sim \mathcal{N}(0, \sigma_x^2)$	$\varepsilon_n \sim \mathcal{N}(0, \sigma_\varepsilon^2)$	$y_n = \theta_0 x_n + \varepsilon_n$

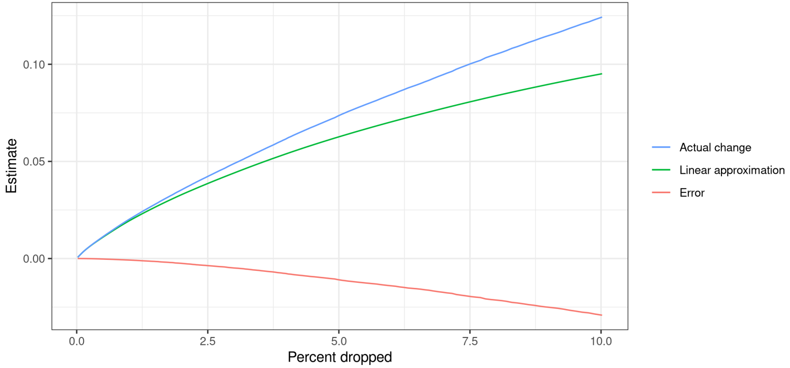


Figure: The actual change, linear approximation to the change, and approximation error.Here, $\sigma_x = 2$, $\sigma_\varepsilon = 1$, and $\theta_0 = 0.5$.

The linear approximation.

For $N = 5,000$ data points, compute the OLS estimator from:

Regressors	Residuals	Responses
$x_n \sim \mathcal{N}(0, \sigma_x^2)$	$\varepsilon_n \sim \mathcal{N}(0, \sigma_\varepsilon^2)$	$y_n = \theta_0 x_n + \varepsilon_n$

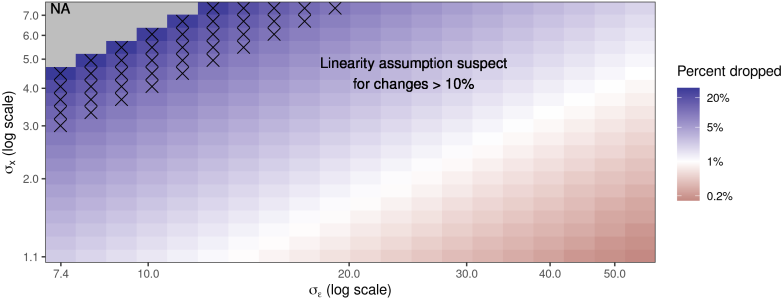


Figure: The approximate perturbation inducing proportion at differing values of σ_x and σ_ε. Red colors indicate datasets whose sign can is predicted to change when dropping less than 1% of datapoints. The grey areas indicate $\hat{\Psi}_\alpha = \mathtt{NA}$, a failure of the linear approximation to locate any way to change the sign.

Conclusions

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is principally determined by the signal to noise ratio, does not disappear asymptotically, and is distinct from (and typically larger than) standard errors.

Conclusion

- You may be concerned if you could reverse your conclusion by removing a $\lfloor \alpha N \rfloor$ datapoints, for some small α.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is principally determined by the signal to noise ratio, does not disappear asymptotically, and is distinct from (and typically larger than) standard errors.
- Robustness to removing a $\lfloor \alpha N \rfloor$ datapoints is easy to check! We can quickly and automatically find an approximate influential set which is accurate for small α.

Links and references

Tamara Broderick, Ryan Giordano, Rachael Meager (alphabetical authors)
"An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?"
https://arxiv.org/abs/2011.14999

See the paper for applications to:

- Hierarchical meta-analysis of microcredit [Meager, 2020]
- Cash transfers randomized controlled trial [Angelucci and De Giorgi, 2009]
- Oregon Medicaid experiment [Finkelstein et al., 2012]
- Expository simulations

`zaminfluence`: `R` package with leave-α-out robustness for OLS and IV estimators
https://github.com/rgiordan/zaminfluence

M. Angelucci and G. De Giorgi. Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508, 2009.

M. Angelucci, D. Karlan, and J. Zinman. Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015.

A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. Newhouse, H. Allen, K. Baicker, and Oregon Health Study Group. The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106, 2012.

R. Giordano, M. I. Jordan, and T. Broderick. A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019.

F. Hampel. *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience, 1986.

R. Meager. Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. *LSE working paper*, 2020.